문헌 클러스터링 결과의 성능 평가 방법에 관한 비교 연구

A Comparative Study on Performance Evaluation of Document Clustering Results

김정하, 이재윤 (연세대학교 대학원 문헌정보학과)

Jung-Ha Kim, Jae Yun Lee
Dept. of Library and Information Science, Graduate School of Yonsei University

자동분류나 정보검색에 활용되는 문헌 클러스터링 결과의 성능을 평가하는 방법에는 여러가지가 있다. 본 논문에서는 제시된 몇 가지 평가방법의 개념과 특징에 대해 알아본다. 학술논문 초록 집합인 KTSET과 신문기사 집합인 KFCM-CL을 대상으로 각각 유사계수를 변화시켜가며 클러스터링한 결과에 대해 각 평가방법을 적용해본 후, 특징과 문제점을 살펴보았다.

1 서론

문헌 또는 용어의 자동분류를 위해 다양한 클러스터링 실험이 수행되고 있다. 분류대상물, 분류자질 선정, 유사성 척도, 클러스터링 알고리즘에 관한 연구가 활발하지만, 평가방법에 대한 연구는 미진한 편이다. 대부분의 문헌 클러스터링 실험에서는 성능평가를 위해 클러스터를 이용한 정보검색 실험을 수행한 다음, 검색 성능을 재현율, 정확률 척도로 측정해왔다. (Burgin 1995; Hearst 1996; Sahami et al. 1998; Zamir 1998). 즉, 검색 성능으로 분류 성능을 대신하고 있는 것이다. 하지만, 재현율과 정확률에 근거한 척도는 '잘 된 클러스터링 결과는 정보검색의 효율성을 증진시킨다'는 것을 전제로 하는, 분류 성능의 간접적인 평가척도이다. 그러므로 자동분류에서는 클러스터링 결과에 대한 직접적인 해석을 위해 형성된 클러스터 자체를 평가할 필요가 있다.

클러스터링 결과 자체를 평가하는 방법으로 몇 가지 제시된 연구가 있으나 이들간 비교는 이루어지지 않았다. 이 연구에서는 제시된 몇 가지 평가방법에 대해 소개하고, 실제 문헌 클러스터링 실험을 통해 각 평가방법의 특징을 살펴보고자 한다.

2 클러스터 타당성

클러스터 타당성(cluster validation)이란 클러스터링 결과를 정량적이고 객관적인 방식으로 평가하는 모든 과정을 일컫는다. 관련된 용어의 정의를 정리하면 다음과 같다.

- o **계층**(Hierarchies) : 데이터에 계층적 클러스터링 알고리즘을 적용하여 얻은 포함된 클러스터의 연속. 덴드로그램으로 표현된다.
- o **분할**(Partitions) : 반복적 분할 클러스터링 알고리즘을 적용하여 얻은 클러스터 집합, 혹은 계층적 클러스터링 알고리즘을 통해 얻은 덴드로그램을 특정 기준치에서 잘랐을 때 형성되는 클러스터 집합.
- o **클러스터**(Clusters) : 하나의 분할을 구성하는 요소.

클러스터의 타당성은 크게 세 부분으로 나눌 수 있다. 계층의 타당성 검증 시에는 계층적 클러스터링 기법을 적용하여 형성된 전체 계층을 고려한다. 분할의 타당성에서는 클러스터링 기법을 적용하여 얻은 분할에서 의미 있는 클러스터는 얼마나 되는가, 클러스터간 구조는 적절한가를 검증한다. 클러스터의 타당성에서 초점이 되는 것은 개개의 클러스터로서, 각 클러스터의 응집도와 분리도를 측정한다.

클러스터 타당성의 평가 기준은 크게 외적 기준과 내적 기준으로 나누어진다.

- o **외적 기준**(External criteria) 클러스터 구조를 이미 마련되어 있는 외부 정보를 이용하여 평가. 원 클러스터의 복원도를 측정.
- o **내적 기준**(Internal criteria) 클러스터링 결과를 외부 정보를 이용하지 않고, 데이터 자체만을 가지고 평가.

문헌 클러스터링에서는 수작업 분류 결과를 외적 기준으로 적용하는 것이 일반적이다.

본 연구에서는 학술논문 초록집합인 KTSET과 신문기사 본문집합인 KFCM-CL을 대상으로 유사계수를 변화시켜가며 클러스터링 실험을 수행하였다. 각 실험의 결과를 분할의 타당성, 클러스터의 타당성 측면에서 평가해보고 각 평가척도의 특징을 분석해보려 한다. 본 실험에서 사용된 클러스터 평가척도 네 가지는 <표 1>과 같다.

<표 1> 실험에 사용된 클러스터 평가 척도

분할의 타당성	외적기준에 의한 척도	① Borko의 χ^2
		② Rand 척도
		③ CSIM
클러스터의 타당성		④ Entropy

3 클러스터링 결과 평가방법

3.1 Borko의 chi-squre 통계치

Borko(1968)는 두 가지 클러스터링 결과를 비교하는 비교적 간단한 방법을 제안하였다. A와 B를 각각 n개의 데이터에 적용한 두 가지 다른 클러스터링 결과라 하고, A는 R개, B는 C개의 클러스터로 다음과 같이 구성될 경우,

$$A = \{a_1, a_2, \ldots a_R\}, \quad B = \{b_1, b_2, \ldots b_C\}$$

<표 2>의 n_{ij}는 클러스터 a_i와 b_j에 동시에 속하는 데이터의 개수를 의미한다.

<표 2> R×C 분할표

	b_1	b_2	⋯	b_C	*total*
a_1	n_{11}	n_{12}	⋯	n_{1C}	$n_{1.}$
a_2	n_{21}	n_{22}	⋯	n_{1C}	$n_{2.}$
⋮	⋮	⋮		⋮	⋮
a_R	n_{R1}	n_{R2}	⋯	n_{RC}	$nR.$
total	$n_{.1}$	$n_{.2}$	⋯	$n_{.C}$	$n_{..}= n$

Borko는 mean contingency(Φ)와 chi-square 통계치(χ^2)를 이용하여 클러스터링 결과간 연관성을 측정하였다. 본 연구에서는 χ^2 통계치를 사용하도록 한다. χ^2 통계치는 두 개의 변인 사이에 서로 유의한 관계가 있는지를 측정한다. R×C 분할표에서 χ^2 통계치는 다음과 같이 정의할 수 있다. O_{ij}는 실제빈도이며, E_{ij}는 기대빈도이다.

$$\chi^2 = \sum_{i=1}^{i=R} \sum_{j=1}^{j=C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

실제빈도 O_{ij}는 n_{ij}이며, ij번째 셀의 기대빈도 E_{ij}는 다음과 같다.

$$E_{ij} = \frac{n_{i.}\, n_{.j}}{n_{..}}$$

χ^2 통계치는 클수록 연관성이 높음을 나타내지만, 표의 크기에 의존한다는 단점이 있다.

3.2 Rand 척도(Rand index)

Rand(1971)는 Borko와는 다른 접근법을 제시하였다. 그는 짝진 쌍의 데이터가 두 가지 다른 클러스터링 기법을 적용한 결과 A와 B에서 각각 동일한 클러스터에 속하는지를 측정하였다. 만약 짝진 데이터 쌍이 두 가지 클러스터링 결과 A, B에서 모두 하나의 클러스터에 속한다면, 이들은 유사하게 취급되었다고 할 수 있다. 만약 짝진 데이터 쌍이 클러스터링 결과 A에서는 하나의 클러스터에 속했지만, B에서는

서로 다른 클러스터에 속했다면, 이들은 유사하게 취급되었다고 할 수 없다. Rand 척도는 2×2 분할표를 사용하여 구할 수 있다. 짝진 데이터 쌍의 모든 경우의 수는 n(n-1)/2개이며, 이들은 다음의 두 범주로 구분된다.

1 : 짝진 쌍이 클러스터링 결과에서 하나의 클러스터에 속한 경우의 수
0 : 짝진 쌍이 클러스터링 결과에서 서로 다른 클러스터에 속한 경우의 수

<표 3> 2×2 분할표

		클러스터링 결과 B: 1	클러스터링 결과 B: 0
클러스터링 결과 A	1	a	b
	0	c	d

Rand 척도에서 두 클러스터링 결과간 유사도는 다음과 같다. Rand 척도는 0과 1사이의 값을 가지며, 값이 클수록 클러스터링 결과간 유사도가 높은 것이다.

$$Rand(C_A, C_B) = \frac{a+d}{a+b+c+d}$$

3.3 CSIM (Cluster SIMilarity)

한승희, 이재윤(1999)은 클러스터링 결과간 유사도를 측정하기 위하여 CSIM을 정의하였다. CSIM은 <표 3>의 분할표에 대해 다이스계수 공식을 적용한 척도이다. 즉, 이 값은 분할표에서 d값을 고려하지 않으면서 a값에 2배의 가중치를 준 값이다.

$$CSIM(C_A, C_B) = \frac{2a}{2a+b+c}$$

이는 결과 A에서 동일한 클러스터에 속한 문헌 쌍이 결과 B에서도 동일한 클러스터에 속할 확률을 의미한다. CSIM 척도도 Rand 척도와 마찬가지로 0과 1사이의 값을 갖는다.

3.4 Entropy

클러스터 내 데이터의 동질성을 측정하는 척도로 정보이론에서 유래된 entropy를 사용하는 경우도 있다.(Han 1998). 클러스터링 알고리즘을 적용하여 얻은 클러스터 C가 있다고 하자. C에 포함된 데이터 중 일부는 미리 부여된 범주 X에 속하고 일부는 범주 Y에 속한다고 하면 C의 entropy는 다음과 같다.

$$Entropy(C) = -p_X \log_2 p_X - p_Y \log_2 p_Y$$

여기서 p_X는 클러스터 C에서 범주 X에 속한 데이터의 비율, p_Y는 C에서 범주 Y에 속한 비율을 의미한다. 실제로 계산을 할 때, 0log0은 0으로 간주한다.

entropy가 0이라는 것은 클러스터 C의 모든 데이터가 동질하다는 것을 뜻한다. 따라서, entropy값이 0에 가까울수록 클러스터링 결과가 우수하다고 할 수 있다.

한 클러스터에 모인 데이터들의 외적 기준에 의한 소속 범주가 n가지라면, entropy를 구하는 식은 다음과 같이 정의할 수 있다.

$$Entropy(C) = \sum_{i=1}^{n} - p_i \log_2 p_i$$

클러스터링 결과 성능 평가 척도로 entropy를 이용하는 경우에는 개별 클러스터의 크기 차이를 반영하도록 각 클러스터 entropy의 가중평균을 산출한다.

4 실험 1 : KTSET 적용

4.1 문헌집단

KTSET 1.0 1,053건 중 소속문헌의 수가 6이상인 14개 범주에 속한 189건의 문서를 대상으로 실험하였다. 14개 범주의 평균 문헌 수는 13.5이며, 범주당 최소문헌 수는 6건, 최대문헌 수는 39건이다. 한국어 형태소 분석기 HAM을 사용하여 제목과 한글초록을 대상으로 자동색인한 결과 생성된 색인어의 수는 총 3061개였다. 이를 이용하여 189×3061 문헌-용어 행렬을 구축하였다.

<표 4> 실험대상 유사계수 14개

유사계수		약칭
가중치 계수	Block	blo
	Cosine	cos
	Euclidean	euc
	Pearson correlation	pea
이진 계수	Dice	dic
	Jaccard	jac
	Kulczynski 2	kul2
	Ochiai	och
	Phi 4-point correlation	phi
	Russel and Rao	R&R
	Sokal and Sneath 2	S&S2
	Sokal and Sneath 4	S&S4
	Sokal and Sneath 5	S&S5
	Yule's Q	yulq
	Yule's Y	yuly

4.1 유사계수

클러스터링 기법 중 완전연결 기법에 15가지 유사계수를 적용한 후, 네 가지 평가척도를 사용하여 이들간 성능을 비교하려 한다. 실험의 대상이 된 유사계수들은 가중치계수 4개와 이진계수 11개이다. <표 4>는 실험의 대상이 된 유사계수와 그 약칭을 나타낸 것이다.

4.2 결과

<표 5>는 15가지 유사계수를 이용하여 얻은 클러스터링 결과에 3장에서 언급한 네 가지 평가방법을 적용한 결과를 요약한 것이다. dic, jac 계수와 S&S2, yulq와 yuly는 클러스터링 결과 동일한 클러스터를 형성하였다. 이를 평가방법별로 나누어 기법간 성능의 우선순위를 살펴보면 <표 6>과 같다.

가중치계수 중 pea, cos과 이진계수 중 kul2의 성능이 우수하게 나타났으며, 가중치계수 중 거리계수인 blo, euc의 성능은 좋지 않았다.

<표 5> KTSET 189건의 클러스터링 결과 평가

유사계수	Borko	Rand	CSIM	entropy
blo	460.266	0.558	0.247	2.596
cos	820.740	0.896	0.483	1.556
euc	404.307	0.437	0.212	2.770
pea	891.189	0.911	0.523	1.464
dic	691.458	0.863	0.339	1.880
jac	691.458	0.863	0.339	1.880
kul2	853.379	0.876	0.396	1.716
och	680.859	0.872	0.357	1.851
phi	699.651	0.879	0.339	1.802
R&R	598.478	0.883	0.319	1.785
S&S2	691.458	0.863	0.339	1.880
S&S4	743.215	0.872	0.397	1.765
S&S5	792.234	0.881	0.393	1.646
yulq	802.115	0.876	0.346	1.737
yuly	802.115	0.876	0.346	1.737

5 실험 2 : 신문기사 적용

5.1 문헌집단

실험 2에서는 문헌집단을 바꾸어 실험하였다. 실험대상이 된 문헌집단은 연세대학교 문헌정보학과에서 구축한 KFCM-CL 집합 중 소속 문헌이 13개 이상인 11개 범주에 속한 신문기사 207건이다. 11개 범주의 평균 문헌 수는 20.7건, 범주당 최대문헌 수는 30건이다. HAM을 이용한 자동색인 후, cf가 1인 용어는 제외시켰으며, 나머지 색인어를 이용해 207×4296 문헌-용어 행렬을 구축하였다.

5.2 결과

<표 7>은 신문기사 문헌집단을 대상으로 15가지 유사계수를 적용하여 얻은 클러스터링 결과에 네 가지 평가방법을 적용한 결과를 요약한 것이다. 실험 2에서도 마찬가지로 dic, jac 계수와 S&S2, yulq와 yuly는 클러스터링 결과

<표 6> 평가방법별 유사계수의 성능 우선순위 - KTSET 189건의 경우

Borko	pea > kul2 > cos > {yulq, yuly} > S&S5 > S&S4 > phi > {dic, jac, S&S2} > och > R&R > blo > euc
Rand	pea > cos > R&R > S&S5 > phi > kul2 = {yulq, yuly} > och = S&S4 > {dic, jac, S&S2} > blo > euc
CSIM	pea > cos > S&S4 > kul2 > S&S5 > och > {yulq, yuly} > {dic, jac, S&S2} = phi > R&R > blo > euc
entropy	pea > cos > S&S5 > kul2 > {yulq, yuly} > S&S4 > R&R > phi > och > {dic, jac, S&S2} > blo > euc

<표 7> KFCM-CL 207건의 클러스터링 결과 평가

유사계수	Borko	Rand	CSIM	entropy
blo	401.077	0.409	0.202	2.668
cos	1182.568	0.924	0.634	0.921
euc	259.223	0.380	0.185	2.831
pea	1256.778	0.939	0.697	0.791
dic	1149.320	0.877	0.485	1.055
jac	1149.320	0.877	0.485	1.055
kul2	977.929	0.850	0.434	1.325
och	1009.224	0.878	0.461	1.180
phi	1278.462	0.932	0.655	0.789
R&R	1032.763	0.895	0.509	1.159
S&S2	1149.320	0.877	0.485	1.055
S&S4	1156.485	0.916	0.582	0.909
S&S5	1105.290	0.855	0.442	1.193
yulq	1139.087	0.909	0.536	0.976
yuly	1139.087	0.909	0.536	0.976

동일한 클러스터를 형성하였다. 이를 평가방법별로 나누어 기법간 성능의 우선순위를 살펴보면 <표 8>과 같다.

가중치계수인 pea, cos의 성능은 여전히 우수하게 나타났으며, 이진계수 중에서는 phi 계수의 성능이 우수하였다. blo, euc, kul2의 성능은 저조하게 나타났다.

6 논의

다른 계수에 비해 현저히 성능이 떨어지는 계수(blo, euc)는 두 실험에서 모두 평가방법에 상관없이 가장 낮은 성능을 보였다.

pea는 모든 경우에 1,2위를 차지하여 높은 성능을 보였다.

중간 순위의 경우, 실험 1에서는 평가방법에 따라 R&R과 같이 순위 차이가 다소 큰 경우가 나타났으며, 실험 2에서는 대부분의 경우 평가방법에 따른 순위 차이가 작게 나타났다.

이상에서 평가방법에 따라서 성능의 우열이 다소 다르게 판정되는 것을 알 수 있다. 평가방법 사이의 차이가 어느 정도인지를 알아보기 위해서 피어슨 상관을 구한 결과를 <표 9>에 제시하였다.

<표 9> 평가방법 사이의 Pearson 상관

실험 1

평가방법	Borko	Rand	CSIM	entropy
Borko	1.000			
Rand	0.849	1.000		
CSIM	0.873	0.733	1.000	
entropy	0.927	0.964	0.862	1.000

실험 2

평가방법	Borko	Rand	CSIM	entropy
Borko	1.000			
Rand	0.980	1.000		
CSIM	0.925	0.906	1.000	
entropy	0.994	0.993	0.932	1.000

* entropy는 다른 방법과 달리 값이 낮을수록 성능이 좋음을 뜻하므로 상관 값의 부호를 반전시켰다.

<표 9>에서 보면 전체적으로 실험 1의 경우가 실험 2보다 평가방법 간의 차이가 큼을 알 수 있으며, CSIM과 Rand 사이가 가장 상관이 낮게 나타났다.

이상에서 평가방법에 따라 성능 판정의 차이가 나타남을 알 수가 있었다. 각 평가방법의 특징과 문제점을 더욱 면밀히 분석하기 위하여 <표 10>과 같이 네 가지 형태의 분할을 임의로 생성한 후, 평가의 외적 기준이 되는 클러스터를 변경해가며 결과를 살펴보았다.

p1 기준에서 Borko, Rand, entropy는 p2과 p4의 성능차이를 구별하지 못한다.

Borko는 p3 기준에서 p2보다 p4를 좋게 판정한다. 그런데 p2가 p3과 같아지기 위해서는 문헌 하나(d4)의 소속만 옮기면 되지만 p4가

<표 10> 임의로 생성한 분할

	p1	p2	p3	p4
c1	d1 d2	d1 d2 d3 d4	d1 d2 d3	d1 d2
c2	d3 d4	d5 d6	d4 d5 d6	d3 d5
c3	d5 d6			d4 d6

p는 분할, c는 클러스터, d는 문헌

<표 8> 평가방법별 유사계수의 성능 우선순위 - KFCM-CL 207건의 경우

Borko	phi > pea > cos > S&S4 > {dic, jac, S&S2} > {yulq, yuly} > S&S5 > R&R > och > kul2 > blo > euc
Rand	pea > phi > cos > S&S4 > {yulq, yuly} > R&R > och > {dic, jac, S&S2} > S&S5 > kul2 > blo > euc
CSIM	pea > phi > cos > S&S4 > {yulq, yuly} > R&R > {dic, jac, S&S2} > och > S&S5 > kul2 > blo > euc
entropy	phi > pea > S&S4 > cos > {yulq, yuly} > {dic, jac, S&S2} > R&R > och > S&S5 > kul2 > blo > euc

<표 11> 평가방법을 적용한 결과

p1 기준	Borko	Rand	CSIM	entropy
p2	6.000	0.733	0.600	0.667
p3	4.000	0.667	0.444	0.918
p4	6.000	0.733	0.333	0.667

p2 기준	Borko	Rand	CSIM	entropy
p1	6.000	0.733	0.600	0.000
p3	3.000	0.667	0.615	0.459
p4	1.500	0.467	0.200	0.667

p3 기준	Borko	Rand	CSIM	entropy
p1	4.000	0.667	0.444	0.333
p2	3.000	0.667	0.615	0.541
p4	4.000	0.667	0.444	0.333

p4 기준	Borko	Rand	CSIM	entropy
p1	6.000	0.733	0.333	0.667
p2	1.500	0.667	0.200	1.333
p3	4.000	0.667	0.444	0.918

p3와 같아지기 위해서는 두 문헌(d3, d5)의 소속을 옮겨야 하므로 옳은 판정으로 보기 어렵다. p3와 p2를 비교할 때에는 <표 2>의 R×C 분할표의 크기가 2×2=4 이지만, p3와 p4간의 비교에서는 2×3=6 이기 때문에 Borko의 경우 p4를 p2보다 과대평가한 것이다. 이와 같이 Borko의 카이제곱 통계치는 표의 크기에 영향을 받기 때문에 클러스터의 수가 고정되지 않으면 절대적인 평가 척도로 이용하기 어렵다.

Rand 척도는 p3을 기준으로 하였을 때, p1과 p2를 구분하지 못한다. <표 3>의 분할표를 아래와 같이 구성해보면 a+d값이 같은 경우 판별능력 없다는 단점이 있다.

p3 기준 p1 평가

	1	0
1	(2)	1
0	4	(8)

p3 기준 p2 평가

	1	0
1	(4)	3
0	2	(6)

entropy는 Borko와 마찬가지로 생성된 클러스터의 수가 많을수록 성능이 좋은 쪽으로 판정할 여지가 높다. p2 기준 평가에서 entropy는 기준 클러스터가 (1, 2, 3, 4)인 경우, 형성된 클러스터가 (1, 2), (3, 4)가 되면 0.0값을 갖는다. 통합되어야 하는 클러스터가 분리되었을 때의 문제점을 반영하지 못한다는 단점이 있다. 또한 entropy는 두 분할 중 어느 것을 기준으로 보는가에 따라 값이 달라진다. 즉, entropy는 분류정확률만 다루고 분류재현율은 무시하는 척도이다.

반면, CSIM은 분류정확률과 분류재현율의 관점에서 볼 때 범주화 평가에서 흔히 쓰이는 F척도와 같다. 여러 가지 평가척도 중 CSIM이 가장 무난한 척도로 판단된다.

7 결론

지금까지 제시된 몇 가지 평가방법의 개념과 특징에 대해 알아보고, 실험을 통해 각 평가방법간에는 어떠한 관계가 있는지 살펴보았다. 평가방법의 특징이 상이하므로 클러스터링 실험을 수행할 때에는 실험의 성격을 검토하여 알맞은 평가방법을 적용하여야 할 것이다. 앞으로 다양한 클러스터링 상황에 대해서 각 평가방법의 이론적인 문제점을 검증하는 실험이 필요하다.

참고문헌

한승희, 이재윤. 1999. 문헌 클러스터링을 위한 유사계수간의 연관성 측정. 제6회 한국정보관리학회 학술대회 논문집, 25-28.

Borko, H. et al. 1968. On-line information retrieval using associative indexing. RADC-TR-68-100. AD670195. Califonia : System Develop. Corp. Quated in Anderberg M.R. *Cluster analysis for applications.* (New York : Academic Press, 1973), 204-207.

Burgin, R. 1995. The retrieval effectiveness of five clustering algorithms as a function of indexing exhaustivity. *JASIS*, 46(8) : 562-572.

Han, Eui-Hong et al. 1998. WebACE: a web agent for document categorization and Exploration. Proceedings of the 2nd International Conference on Autonomous Agents.

Hearst, M. A. and Pedersen, J. O. 1996. Reexamining the cluster hypothesis : Scatter/Gather on retrieval results. *ACM SIGIR '96.* 76-84.

Rand, W.M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association,* 66 : 846-850.

Sahami, M. et al. 1998. SONIA : a service for organizing networked information autonomously. Digital Libraries 98.

Zamir, O. and Etzioni, O. 1998. Web document clustering : a feasiility demonstration. *ACM SIGIR '98.* 46-54.